Journal of Machine Learning Research 18 (2017) 1-26 Submitted 11/15; Revised 10/16; Published 7/17

Statistical Inference with Unnormalized Discrete Models and Localized Homogeneous Divergences

Takashi Takenouchi* TTAKASHI@FUN.AC.JP
Future University Hakodate
RIKEN Center for Advanced Intelligence Project
116-2 Kamedanakano, Hakodate, Hokkaido, 040-8655, Japan

Takafumi Kanamori KANAMORI@IS.NAGOYA-U.AC.JP
Department of Computing and Software Systems, Nagoya University
RIKEN Center for Advanced Intelligence Project
Furocho, Chikusaku, Nagoya 464-8603, Japan

Editor: Guy Lebanon

Abstract

In this paper, we focus on parameters estimation of probabilistic models in discrete space. A naive calculation of the normalization constant of the probabilistic model on discrete space is often infeasible and statistical inference based on such probabilistic models has difficulty. In this paper, we propose a novel estimator for probabilistic models on discrete space, which is derived from an empirically localized homogeneous divergence. The idea of the empirical localization makes it possible to ignore an unobserved domain on sample space, and the homogeneous divergence is a discrepancy measure between two positive measures and has a weak coincidence axiom. The proposed estimator can be constructed without calculating the normalization constant and is asymptotically consistent and Fisher efficient. We investigate statistical properties of the proposed estimator and reveal a relationship between the empirically localized homogeneous divergence and a mixture of the α-divergence. The α-divergence is a non-homogeneous discrepancy measure that is frequently discussed in the context of information geometry. Using the relationship, we also propose an asymptotically consistent estimator of the normalization constant. Experiments showed that the proposed estimator comparably performs to the maximum likelihood estimator but with drastically lower computational cost.

Keywords: unnormalized model, homogeneous divergence, empirical localization, discrete model

1. Introduction

In the fields of machine learning and pattern recognition, probabilistic models on discrete space are useful for classification tasks or modeling discrete phenomena, so estimating parameters of probabilistic models on discrete space is a widely studied and important challenge. For example, the Boltzmann machine (with hidden variables) (Hinton and Sejnowski, 1986; Ackley et al., 1985; Amari et al., 1992) is a widely used probabilistic model to represent binary variables, and the restricted Boltzmann machine (RBM) is especially attracting increasing attention in the context of deep learning (Hinton, 2010; Hinton and Salakhutdinov, 2012). Training of probabilistic models on discrete space, i.e., estimation of parameters, is

usually done by using the maximum likelihood estimation (MLE). An explicit expression of the MLE cannot generally be obtained, so the gradient-based optimization is usually used. A difficulty of the gradient-based optimization for such models comes from the calculation of the normalization constant. The calculation of the gradient requires calculation of the normalization constant in each step of the optimization and its computational cost sometimes increases exponentially. To tackle the problem of computational cost, various kinds of approximation methods have been proposed. One approach tries to approximate the probabilistic model (or expectation with the model) by using a tractable model or sampling techniques. The mean-field approximation is a popular method to approximate the model by assuming independence among variables (Opper and Saad, 2001). The contrastive divergence (Hinton, 2002) avoids the exponential order calculation using the Markov Chain Monte Carlo (MCMC) sampling: the method approximates the expectation using samples obtained by the MCMC that runs few steps from the empirical distribution with a transition matrix defined with a current model. Another approach is estimation based on an unnormalized model, which does not include the normalization constant and so is not necessarily a probability. In the literature of parameters estimation of probabilistic models for continuous variables, Hyvärinen (2005) used a score function which is a gradient of log-density with respect to the data vector rather than parameters. This approach makes it possible to estimate parameters without calculating the normalization term by focusing on the shape of the density function. Hyvärinen (2007) extended the method to discrete variables. This method uses information of a "neighbor" by contrasting its probability with that of a flipped variable. Gutmann and Hirayama (2012) extended the above framework and developed an approximated estimator on the basis of the Bregman divergence. Dawid et al. (2012) proposed a generalized local scoring rules on discrete sample spaces and theoretically discussed a class of scoring rules for appropriate estimation. Takenouchi (2015) focused on a specific class of models on discrete space and constructed an asymptotically consistent estimator using the Itakura-Saito distance.

In this paper, we propose a novel method of parameter estimation for probabilistic models on discrete space. The proposed estimator can be constructed without calculation of the normalization constant by utilizing the unnormalized model and has suitable statistical properties such as consistency and efficiency.[1] The proposed estimator is derived from minimization of a risk function defined by a combination of two ideas: the homogeneous divergence and an empirical localization. The conventional divergence measure that has a coincidence axiom ensures that the divergence is zero if and only if two positive measures are equal. On the other hand, the homogeneous divergence follows a weak coincidence axiom and is zero if and only if two positive measures have a proportional relationship. By virtue of the property, the homogeneous divergence that has the weak coincidence axiom can ignore the normalization constant, and the proposed estimator does not need to calculate the normalization constant. The empirical localization is a method of transformation of the unnormalized model and is an extension of the geometric mean of the model and the

1. A short version of this article was published as a conference paper (Takenouchi and Kanamori, 2015). In this paper, we show the possibility of generalization of the basic framework shown in (Takenouchi and Kanamori, 2015) and add some new theoretical results, including estimator for the normalization constant. Also we add experiments with various kinds of models and settings to assess the validity of the proposed method.

empirical distribution of a given dataset. By using the empirical localization, we can restrict the domain of the estimation problem to that of observed examples and omit calculations associated with the domain of unobserved examples. This restriction of the domain can drastically reduce computational cost. The derived risk function is convex for various kinds of models including the higher order Boltzmann machine and so is easy to optimize. Also, we investigate statistical properties of the proposed estimator and reveal that the proposed estimator has the same asymptotic efficiency as the MLE. The proposed risk function is closely related with the α-divergence (Amari and Nagaoka, 2000), which induces a Fisher efficient estimator for discrete probabilistic models. The α-divergence is not homogeneous divergence, and by utilizing the relationship, we also developed a feasible estimator of the normalization constant.

Basic settings are described in Section 2, and we introduce the homogeneous divergence and the idea of empirical localization of the unnormalized model in Section 3. We describe a novel estimator for probabilistic models on discrete spaces and investigate its statistical properties of the proposed estimator in Section 4. In Section 5, we discuss a characterization of the risk function of the proposed estimator with the α-divergence and construct a feasible estimator of the normalization constant. We numerically verify performance of the proposed estimator by using synthetic datasets in Section 6.

2. Settings

Let $\mathcal{X}$ be a discrete space such as $\{+1, -1\}^d$ or the set of natural numbers. The bracket $\langle f \rangle$ for a real-valued function f on $\mathcal{X}$ denotes the sum of $f(\boldsymbol{x})$ over $\mathcal{X}$ i.e., $\langle f \rangle = \sum_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$. We define $\mathcal{M}$ and $\mathcal{P}$ as the set of all non-negative functions and that of all probability functions on $\mathcal{X}$,

$$\mathcal{M} = \{ f : \mathcal{X} \to \mathbb{R}_{\geq 0} \,|\, \langle f \rangle < \infty \}, \ \mathcal{P} = \{ f \in \mathcal{M} \,|\, \langle f \rangle = 1 \}$$

where $\mathbb{R}_{\geq 0}$ is the set of all non-negative real numbers.

In this paper, we focus on parameter estimation of a probabilistic model $\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$ on $\mathcal{X}$ that is written as

$$\bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{q_{\boldsymbol{\theta}}(\boldsymbol{x})}{Z_{\boldsymbol{\theta}}} \tag{1}$$

where $\boldsymbol{\theta}$ is an m-dimensional vector of parameters, $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ is an unnormalized model in $\mathcal{M}$ and $Z_{\boldsymbol{\theta}} = \langle q_{\boldsymbol{\theta}} \rangle$ is the normalization constant. The equality $\langle q_{\boldsymbol{\theta}} \rangle = \sum_{\boldsymbol{x} \in \mathcal{X}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1$ does not necessarily hold for unnormalized models, and the normalization constant $Z_{\boldsymbol{\theta}}$ typically requires a high computational cost. We thus need a device to circumvent the computation of Z_θ in statistical inferences. Throughout the paper, we assume without loss of generality that the unnormalized model $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ can be written as

$$q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\psi_{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{2}$$

where $\psi_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a function on $\mathcal{X}$ with the parameter $\boldsymbol{\theta}$.

Remark 1 *By setting $\psi_{\boldsymbol{\theta}}(\boldsymbol{x})$ as $\psi_{\boldsymbol{\theta}}(\boldsymbol{x}) - \log Z_{\boldsymbol{\theta}}$, the normalized model* (1) *can be written as* (2).

Example 1 *The Bernoulli distribution on $\mathcal{X} = \{+1, -1\}$ is the simplest example of the probabilistic model* (1) *using the function $\psi_\theta(x) = \theta x$ for $x \in \mathcal{X}$ and $\theta \in \mathbb{R}$.*

Example 2 *For the parameter $\theta \in \mathbb{R}$, the function $\psi_\theta(x) = x\theta - \log x!$ leads to an unnormalized model for the Poisson distribution on $\mathcal{X} = \{0, 1, 2, \ldots\}$ with mean e^θ.*

Example 3 *Using a function $\psi_{\boldsymbol{\theta},k}(\boldsymbol{x}) = (x_1, \ldots, x_d, x_1x_2, \ldots, x_{d-1}x_d, x_1x_2x_3, \ldots)\boldsymbol{\theta}$ in which monomials of degree up to k appear, we can define a k-th order Boltzmann machine (Hinton and Sejnowski, 1986; Sejnowski, 1986).*

Example 4 *Let $\boldsymbol{x}_o \in \{+1, -1\}^{d_1}$ and $\boldsymbol{x}_h \in \{+1, -1\}^{d_2}$ be an observed vector and hidden vector, respectively, and $\boldsymbol{x} = \left(\boldsymbol{x}_o^T, \boldsymbol{x}_h^T\right) \in \{+1, -1\}^{d_1+d_2}$ where T indicates the transpose, be a concatenated vector. The Boltzmann machine with hidden variables is defined as $q_{h,\boldsymbol{\theta}}(\boldsymbol{x}_o) = \exp(\psi_{h,\boldsymbol{\theta}}(\boldsymbol{x}_o))$ where the function $\psi_{h,\boldsymbol{\theta}}(\boldsymbol{x}_o)$ is*

$$\psi_{h,\boldsymbol{\theta}}(\boldsymbol{x}_o) = \log \sum_{\boldsymbol{x}_h} \exp(\psi_{\boldsymbol{\theta},2}(\boldsymbol{x})) \tag{3}$$

and $\sum_{\boldsymbol{x}_h}$ denotes the summation with respect to the hidden variable $\boldsymbol{x}_h$.

Example 5 *Let $\boldsymbol{x}_o \in \{+1, -1\}^{d_1}$ and $\boldsymbol{x}_h \in \{+1, -1\}^{d_2}$ be the observed vector and hidden vector, respectively. The restricted Boltzmann machine is written as*

$$\begin{aligned} q_{\boldsymbol{\theta}}(\boldsymbol{x}_o) &= \sum_{\boldsymbol{x}_h} \exp\left(\boldsymbol{\theta}_o^T \boldsymbol{x}_o + \boldsymbol{\theta}_h^T \boldsymbol{x}_h + \boldsymbol{x}_h^T \boldsymbol{\theta}_{h,o} \boldsymbol{x}_o\right) \\ &= \exp\{\boldsymbol{\theta}_o^T \boldsymbol{x}_o\} \prod_{k=1}^{d_2} \left\{ e^{(\boldsymbol{\theta}_h + \boldsymbol{\theta}_{h,o}\boldsymbol{x}_o)_k} + e^{-(\boldsymbol{\theta}_h + \boldsymbol{\theta}_{h,o}\boldsymbol{x}_o)_k} \right\}, \end{aligned} \tag{4}$$

where $\boldsymbol{\theta}_o \in R^{d_1}, \boldsymbol{\theta}_h \in R^{d_2}$ are vectors of parameters, and $\boldsymbol{\theta}_{h,o} \in R^{d_2 \times d_1}$ is a matrix of parameters. The index k in the above equation denotes the k-th element of the vector. Note that some parameters in the Boltzmann machine with hidden variables (3) *are restricted to* 0.

Suppose that a dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ generated from an underlying distribution $p(\boldsymbol{x})$ is observed. Let $\mathcal{Z}$ be a set of all patterns in the dataset $\mathcal{D}$. An empirical distribution $\tilde{p}(\boldsymbol{x})$ associated with the dataset $\mathcal{D}$ is defined as

$$\tilde{p}(\boldsymbol{x}) = \begin{cases} \frac{n_{\boldsymbol{x}}}{n} & \boldsymbol{x} \in \mathcal{Z}, \\ 0 & \text{otherwise}, \end{cases}$$

where $n_{\boldsymbol{x}}$ is the number of patterns $\boldsymbol{x}$ in the dataset $\mathcal{D}$. To estimate the parameter $\boldsymbol{\theta}$ of the probabilistic model $\bar{q}_{\boldsymbol{\theta}}$, the MLE defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{mle}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

is frequently used, where

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

is the log-likelihood of the parameter $\boldsymbol{\theta}$ using the normalized model $\bar{q}_{\boldsymbol{\theta}}$. The MLE is asymptotically consistent and efficient. However, the optimization of the log-likelihood function can be computationally demanding for normalized models on huge discrete sample spaces. Indeed, the gradient of $L(\boldsymbol{\theta})$ includes $\langle \tilde{p}\psi'_{\boldsymbol{\theta}} \rangle - \langle \bar{q}_{\boldsymbol{\theta}}\psi'_{\boldsymbol{\theta}} \rangle$, where $\psi'_{\boldsymbol{\theta}} = \frac{\partial \psi_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}$. While the first term $\langle \tilde{p}\psi'_{\boldsymbol{\theta}} \rangle$ is the empirical mean that is easily calculated, the second term $\langle \bar{q}_{\boldsymbol{\theta}}\psi'_{\boldsymbol{\theta}} \rangle$ requires 2^d times summation for $\mathcal{X} = \{+1, -1\}^d$. Therefore, the gradient-based optimization is computationally infeasible when d is large. To resolve this problem, we propose a novel estimator using ideas of a homogeneous divergence and an empirical localization in the following section.

3. Homogeneous Divergences for Statistical Inference

A divergence is an extension of the squared distance and is often used in statistical inference. The formal definition of the divergence $D(f, g)$ is a non-negative valued function on $\mathcal{M} \times \mathcal{M}$ or $\mathcal{P} \times \mathcal{P}$ such that $D(f, f) = 0$ holds for arbitrary f. Many popular divergences such as the Kullback-Leilber (KL) divergence defined on $\mathcal{P} \times \mathcal{P}$ enjoy the *coincidence axiom*, i.e., $D(f, g) = 0$ leads to $f = g$. The parameter in the statistical model $\bar{q}_{\boldsymbol{\theta}}$ is estimated by minimizing the divergence $D(\tilde{p}, \bar{q}_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$.

In statistical inference using unnormalized models, the coincidence axiom of the divergence is not suitable since the probability and the unnormalized model generally do not exactly match. Our purpose is to estimate the underlying distribution up to a constant factor using unnormalized models. Hence, a divergence that has the property of the weak coincidence axiom, i.e., $D(f, g) = 0$ if and only if $g = cf$ for some $c > 0$, is a good candidate. For a class of divergences having the weak coincidence axiom, we focus on homogeneous divergences that satisfy the equality $D(f, g) = D(f, cg)$ for any $f, g \in \mathcal{M}$ and any $c > 0$.

A representative of homogeneous divergences is the pseudo-spherical (PS) divergence (Good, 1971), or in other words, γ–divergence (Fujisawa and Eguchi, 2008), that is defined from the (reverse) Hölder inequality. For a positive constant γ and all non-negative functions f, g in $\mathcal{M}$, the Hölder inequality

$$\left\langle f^{1+\gamma} \right\rangle^{\frac{1}{1+\gamma}} \left\langle g^{1+\gamma} \right\rangle^{\frac{\gamma}{1+\gamma}} \geq \langle fg^{\gamma} \rangle \quad (\gamma > 0)$$

holds. Also for a negative constant $\gamma(\neq -1)$, the reverse Hölder inequality

$$\left\langle f^{1+\gamma} \right\rangle^{\frac{1}{1+\gamma}} \left\langle g^{1+\gamma} \right\rangle^{\frac{\gamma}{1+\gamma}} \leq \langle fg^{\gamma} \rangle \quad (\gamma < 0, \gamma \neq -1)$$

holds. The above inequalities become an equality if and only if f and g are linearly dependent. From the standard and reverse Hölder inequalities, the PS-divergence $D_\gamma(f, g)$ for $f, g \in \mathcal{M}$ is defined as

$$D_\gamma(f, g) = \mathrm{sgn}(\gamma) \left\{ \frac{1}{1+\gamma} \log \left\langle f^{1+\gamma} \right\rangle + \frac{\gamma}{1+\gamma} \log \left\langle g^{1+\gamma} \right\rangle - \log \langle fg^{\gamma} \rangle \right\} \quad (\gamma \neq 0, -1), \quad (5)$$

where the sign function $\mathrm{sgn}(\gamma)$ takes 1 for $\gamma > 0$ and -1 for $\gamma < 0$. The PS divergence is homogeneous, and the Hölder inequalities ensure the non-negativity and the weak coincidence axiom of the PS-divergence. A scaled PS-divergence converges to the extended KL-divergence defined on $\mathcal{M} \times \mathcal{M}$, as $\gamma \to 0$. Fujisawa and Eguchi (2008) used the PS-divergence with $\gamma > 0$ to obtain a robust estimator in parametric statistical inference.

The PS-divergence $D_\gamma(f, g)$ from the empirical distribution $f = \tilde{p}$ to the unnormalized model $g = q_{\boldsymbol{\theta}}$ is written as

$$D_\gamma(\tilde{p}, q_{\boldsymbol{\theta}}) = Const + \mathrm{sgn}(\gamma)\frac{\gamma}{1+\gamma}\log\sum_{\boldsymbol{x}\in\mathcal{X}} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1+\gamma} - \mathrm{sgn}(\gamma)\log\sum_{\boldsymbol{x}\in\mathcal{Z}}\frac{n_{\boldsymbol{x}}}{n}q_{\boldsymbol{\theta}}(\boldsymbol{x})^{\gamma} \tag{6}$$

and computation of the second term is infeasible in our setup. To avoid such expensive computation, some approximation techniques have been proposed such as the MCMC. Here, we employ a new trick called "empirical localization." The empirical localization is a way to transform the model $q_{\boldsymbol{\theta}}$, and its idea is to localize the domain $\mathcal{X}$ of $q_{\boldsymbol{\theta}}$ to $\mathcal{Z}$ by considering an extension of the geometric mean with the empirical distribution $\tilde{p}$. The empirical localization of $q_{\boldsymbol{\theta}}$ with $\tilde{p}$ is defined by

$$\tilde{p}(\boldsymbol{x})^\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} = \begin{cases} \left(\frac{n_{\boldsymbol{x}}}{n}\right)^\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} & \boldsymbol{x}\in\mathcal{Z} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where α is a non-negative real number. The empirical localization forces the transformed function to be zero on a domain $\mathcal{X}\setminus\mathcal{Z}$ of unobserved points, satisfying $\tilde{p}(\boldsymbol{x}) = 0$. By using the localization trick, the total sum $\langle q_{\boldsymbol{\theta}}^{\gamma+1}\rangle$ is replaced with a quantity similar to the empirical mean,

$$\left\langle \tilde{p}^\alpha q_{\boldsymbol{\theta}}^{1-\alpha}\right\rangle = \sum_{\boldsymbol{x}\in\mathcal{Z}}\left(\frac{n_{\boldsymbol{x}}}{n}\right)^\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}. \tag{8}$$

The above quantity is easy to compute unless the sample size is extremely large. If the sample size is large, a sub-sampling technique for obtaining an approximate of the empirical mean can be used. Note that the empirical localization (7) is not defined when the value of α is negative and the empirical distribution $\tilde{p}(\boldsymbol{x})$ is zero at some points. However, we can formally employ the trick of empirical localization (8) in such cases by ignoring a domain of unobserved points. Thus, in the following, we assume that α is a non-zero real number.

Remark 2 *The summation in* (8) *is defined on* $\mathcal{Z}$ *and is then computable even when* $\alpha < 0$. *Also the summation includes only* $\mathcal{Z}(\leq n)$ *terms, and its computational cost is* $\mathcal{O}(n)$.

In addition, for convenience of notations, we define an e-mixture model $r_{\alpha,\boldsymbol{\theta}}$ ($\tilde{r}_{\alpha,\boldsymbol{\theta}}$) of the unnormalized model (2) and p ($\tilde{p}$) with ratio $\alpha\in\mathbb{R}$, as follows (Amari and Nagaoka, 2000).

$$r_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x})^\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}}{\left\langle p^\alpha q_{\boldsymbol{\theta}}^{1-\alpha}\right\rangle}, \quad \tilde{r}_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\tilde{p}(\boldsymbol{x})^\alpha q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha}}{\left\langle \tilde{p}^\alpha q_{\boldsymbol{\theta}}^{1-\alpha}\right\rangle}.$$

Note that $\tilde{r}_{\alpha,\boldsymbol{\theta}}$ is a normalized version of the empirical localization (7). If p or q takes zero, the parameter α is properly restricted.

Remark 3 *We observe that* $r_{0,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{r}_{0,\boldsymbol{\theta}}(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}}(\boldsymbol{x})$, $r_{1,\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x})$, $\tilde{r}_{1,\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})$. *Also, if* $p(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$, $r_{\alpha,\boldsymbol{\theta}_0}(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ *holds for an arbitrary* α.

To use the trick of empirical localization, we define the *localized PS-divergence* $S_{\alpha,\alpha',\gamma}(p,q)$ for the probability distribution $p \in \mathcal{P}$ and the unnormalized model $q \in \mathcal{M}$ by considering the homogeneous divergence between $f = (p^\alpha q^{1-\alpha})^{\frac{1}{1+\gamma}}$ and $g = (p^{\alpha'} q^{1-\alpha'})^{\frac{1}{1+\gamma}}$ where α, α' are two distinct real numbers:

$$\begin{aligned} S_{\alpha,\alpha',\gamma}(p,q) &= D_\gamma((p^\alpha q^{1-\alpha})^{1/(1+\gamma)}, (p^{\alpha'} q^{1-\alpha'})^{1/(1+\gamma)}) \\ &= \mathrm{sgn}(\gamma) \left\{ \frac{1}{1+\gamma} \log \left\langle p^\alpha q^{1-\alpha} \right\rangle + \frac{\gamma}{1+\gamma} \log \langle p^{\alpha'} q^{1-\alpha'} \rangle - \log \left\langle p^{\bar{\alpha}} q^{1-\bar{\alpha}} \right\rangle \right\}, \end{aligned} \tag{9}$$

where $\bar{\alpha} = (\alpha + \gamma\alpha')/(1+\gamma)$. Note that the localized PS-divergence vanishes if and only if $p^\alpha q^{1-\alpha} \propto p^{\alpha'} q^{1-\alpha'}$, i.e., $q \propto p$. Substituting the empirical distribution $\tilde{p}$ into p, the total sum over $\mathcal{X}$ is replaced with a variant of the empirical mean (8) on $\mathcal{Z}$.

Since $S_{\alpha,\alpha',\gamma}(p,q) = S_{\alpha',\alpha,1/\gamma}(p,q)$ holds, we can assume $\alpha > \alpha'$. Also we observe that for $\gamma < -1$

$$S_{\alpha,\alpha',\gamma}(p,q) = \frac{\gamma}{1+\gamma} S_{\alpha, \frac{\alpha+\gamma\alpha'}{1+\gamma}, -1-\gamma}(p,q),$$

and for $-1 < \gamma < 0$

$$S_{\alpha,\alpha',\gamma}(p,q) = \frac{1}{1+\gamma} S_{\alpha', \frac{\alpha+\gamma\alpha'}{1+\gamma}, -\gamma}(p,q),$$

respectively. Hence, we can assume $\gamma > 0$ without loss of generality. In summary, the conditions of the real parameters α, α', γ in $S_{\alpha,\alpha',\gamma}$ are given by

$$0 < \gamma, \quad \alpha > \alpha', \ \alpha \neq 0, \ \alpha' \neq 0, \ \alpha + \gamma\alpha' \neq 0,$$

where the last condition implies $\bar{\alpha} \neq 0$.

Let us consider another aspect of the computational issue of the localized PS-divergence (9). For the probability distribution p and the unnormalized exponential model $q_{\boldsymbol{\theta}}$, we show that the localized PS-divergence $S_{\alpha,\alpha',\gamma}(p, q_{\boldsymbol{\theta}})$ is convex in $\boldsymbol{\theta}$ when the parameters α, α' and γ are properly chosen.

Theorem 4 *Let* $p \in \mathcal{P}$ *be any probability distribution and* $q_{\boldsymbol{\theta}}$ *be the unnormalized exponential model* $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x})\right)$, *where* $\boldsymbol{\phi}(\boldsymbol{x})$ *is a vector-valued function corresponding to the sufficient statistic in the (normalized) exponential model* $\bar{q}_{\boldsymbol{\theta}}$. *When* α, α' *and* γ *satisfy* $\bar{\alpha} = (\alpha + \gamma\alpha')/(1+\gamma) = 1$, *the localized PS-divergence* $S_{\alpha,\alpha',\gamma}(p, q_{\boldsymbol{\theta}})$ *is convex in* $\boldsymbol{\theta}$. *Otherwise, there exist* p *and* $\boldsymbol{\phi}(\boldsymbol{x})$ *such that* $S_{\alpha,\alpha',\gamma}(p, q_{\boldsymbol{\theta}})$ *is not convex in* $\boldsymbol{\theta}$.

The proof of Theorem 4 is found in Appendix A.

When the value of α is negative and the probability p vanishes at some points, the total sum $\left\langle p^\alpha q^{1-\alpha} \right\rangle$ is not formally defined. We can avoid such a situation by setting both α and $\alpha\prime$ to positive values and observe the following proposition indicating that the result in Theorem 4 holds even if both α and α' are assumed to be positive.

Proposition 5 *Even though* α *and* α' *in Theorem 4 are both restricted to positive numbers, we need* $\bar{\alpha} = 1$ *to guarantee the convexity of the localized PS-divergence* $S_{\alpha,\alpha',\gamma}(p, q_{\boldsymbol{\theta}})$ *in* $\boldsymbol{\theta}$ *for any* p *and any* $\boldsymbol{\phi}(\boldsymbol{x})$.

The proof is found in Appendix B.

The localized PS-divergence with $\bar{\alpha} = 1$ characterized by Theorem 4 is denoted as $S_{\alpha,\alpha'}(p,q)$ where

$$S_{\alpha,\alpha'}(p,q) = \frac{1-\alpha'}{\alpha-\alpha'} \log \left\langle p^{\alpha} q^{1-\alpha} \right\rangle + \frac{\alpha-1}{\alpha-\alpha'} \log \langle p^{\alpha'} q^{1-\alpha'} \rangle. \tag{10}$$

for $\alpha > 1 > \alpha' \neq 0$. The parameter α' can be negative if the probability function p does not take zero on $\mathcal{X}$. Clearly, $S_{\alpha,\alpha'}(p,q)$ satisfies the homogeneity and the weak coincidence axiom as well as $S_{\alpha,\alpha',\gamma}(p,q)$.

The generalized Hölder's inequality admits an extension of the localized PS-divergence. For $f_1, \ldots, f_L \in \mathcal{M}$, the generalized Hölder's inequality

$$\left\langle \prod_{\ell=1}^{L} f_{\ell}^{\delta_\ell} \right\rangle \leq \prod_{\ell=1}^{L} \langle f_\ell \rangle^{\delta_\ell} \tag{11}$$

holds, where $\delta_1, \ldots, \delta_L$ satisfy $0 < \delta_\ell$ and $\sum_\ell \delta_\ell = 1$. The inequality becomes an equality if and only if all $f_1, \ldots, f_L$ are proportional to a non-negative function $g \in \mathcal{M}$. For the sake of completeness, the proof of the generalized Hölder's inequality is shown in Appendix C.

Substituting $p^{\alpha_\ell} q^{1-\alpha_\ell}$ into f_ℓ of the generalized Hölder's inequality, we obtain an extension of the localized PS-divergence defined as

$$S_{\boldsymbol{\alpha},\boldsymbol{\delta}}(p,q) = \sum_{\ell=1}^{L} \delta_\ell \log \left\langle p^{\alpha_\ell} q^{1-\alpha_\ell} \right\rangle - \log \left\langle p^{\bar{\alpha}} q^{1-\bar{\alpha}} \right\rangle \tag{12}$$

with $\bar{\alpha} = \sum_\ell \alpha_\ell \delta_\ell$. Theorem 4 can be extended for $S_{\boldsymbol{\alpha},\boldsymbol{\delta}}(p,q)$. We omit the proof since it is straightforward.

4. Estimation with Localized Pseudo-Spherical Divergences

Given the empirical distribution $\tilde{p}$ and the unnormalized model $q_{\boldsymbol{\theta}}$, we define a novel estimator $\hat{\boldsymbol{\theta}}$ with the localized PS-divergence $S_{\alpha,\alpha',\gamma}$ or $S_{\alpha,\alpha'}$:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \operatorname*{argmin}_{\boldsymbol{\theta}} S_{\alpha,\alpha',\gamma}(\tilde{p}, q_{\boldsymbol{\theta}}) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \Big\{ \frac{1}{1+\gamma} \log \sum_{\boldsymbol{x} \in \mathcal{Z}} \Big(\frac{n_{\boldsymbol{x}}}{n} \Big)^{\alpha} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} + \frac{\gamma}{1+\gamma} \log \sum_{\boldsymbol{x} \in \mathcal{Z}} \Big(\frac{n_{\boldsymbol{x}}}{n} \Big)^{\alpha'} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha'} \\ &\qquad - \log \sum_{\boldsymbol{x} \in \mathcal{Z}} \Big(\frac{n_{\boldsymbol{x}}}{n} \Big)^{\bar{\alpha}} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\bar{\alpha}} \Big\}. \end{aligned} \tag{13}$$

Though the localized PS-divergence plugged-in the empirical distribution is not well-defined when $\alpha' < 0$ and $\tilde{p}(\boldsymbol{x}) = 0$, we can formally define the estimator by restricting the domain $\mathcal{X}$ to the observed set of examples $\mathcal{Z}$, even for such the case, as shown in Remark 2.

Proposition 6 *For the unnormalized model* (2)*, the estimator* (13) *is Fisher consistent.*

Proof We observe

$$\frac{\partial}{\partial \boldsymbol{\theta}} S_{\alpha,\alpha',\gamma}(\bar{q}_{\boldsymbol{\theta}_0}, q_{\boldsymbol{\theta}})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \left(\bar{\alpha} - \frac{\alpha+\gamma\alpha'}{1+\gamma}\right) \langle \bar{q}_{\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \rangle = 0$$

implying the Fisher consistency of $\hat{\boldsymbol{\theta}}$. ■

Example 6 *For the Bernoulli distribution on $\mathcal{X} = \{+1, -1\}$, the estimator* (13) *is equivalent to the MLE, i.e., $\frac{1}{2}\log\frac{n_{+1}}{n_{-1}}$.*

Theorem 7 *Let $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ be the unnormalized model* (2)*, and $\boldsymbol{\theta}_0$ be the true parameter of the underlying distribution $p(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$. Then, the asymptotic distribution of the estimator* (13) *is the normal distribution given as*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1}),$$

where $I(\boldsymbol{\theta}_0)$ is the Fisher information matrix of the normalized model $\bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$.

The proof is shown in Appendix D. The estimator defined from the general localized PS-divergence $S_{\boldsymbol{\alpha},\boldsymbol{\delta}}(\tilde{p}, q_{\boldsymbol{\theta}})$ defined by (12) has the same asymptotic property.

Remark 8 *The asymptotic distribution of* (13) *is equal to that of the MLE, and its variance does not depend on α, α', γ.*

Remark 9 *As shown in Remark 1, the normalized model* (1) *is a special case of the unnormalized model* (2) *and then Theorem 7 holds for the normalized model.*

5. Characterization of $S_{\alpha,\alpha'}$

Let us consider theoretical properties of the localized PS-divergence. In the following subsections, we discuss an influence of selection of α, α' and a characterization of $S_{\alpha,\alpha'}$ defined by (10).

5.1 Influence of Selection of α, α'

We investigate influence of selection of α, α' for the localized PS-divergence $S_{\alpha,\alpha'}$ with a view of the estimating equation. The estimator $\hat{\boldsymbol{\theta}}$ derived from $S_{\alpha,\alpha'}$ satisfies

$$\frac{\partial S_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \propto \left\langle \tilde{r}_{\alpha',\hat{\boldsymbol{\theta}}} \psi'_{\hat{\boldsymbol{\theta}}} \right\rangle - \left\langle \tilde{r}_{\alpha,\hat{\boldsymbol{\theta}}} \psi'_{\hat{\boldsymbol{\theta}}} \right\rangle = 0. \tag{14}$$

which is a moment matching with respect to two distributions $\tilde{r}_{\alpha,\boldsymbol{\theta}}$ and $\tilde{r}_{\alpha',\boldsymbol{\theta}}$ $(\alpha, \alpha' \neq 0, 1)$. On the other hand, the estimating equation of the MLE is written as

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{mle}} \propto \langle \tilde{p} \psi'_{\boldsymbol{\theta}_{mle}} \rangle - \langle \bar{q}_{\boldsymbol{\theta}_{mle}} \psi_{\boldsymbol{\theta}_{mle}} \rangle = \langle \tilde{r}_{1,\boldsymbol{\theta}_{mle}} \psi'_{\boldsymbol{\theta}_{mle}} \rangle - \langle \tilde{r}_{0,\boldsymbol{\theta}_{mle}} \psi'_{\boldsymbol{\theta}_{mle}} \rangle = 0, \tag{15}$$

which is a moment matching with respect to the empirical distribution $\tilde{p} = \tilde{r}_{1,\boldsymbol{\theta}_{mle}}$ and the normalized model $\bar{q}_{\boldsymbol{\theta}} = \tilde{r}_{0,\boldsymbol{\theta}_{mle}}$. While the localized PS-divergence $S_{\alpha,\alpha'}$ is not defined for $(\alpha, \alpha') = (0, 1)$, comparison of (14) with (15) implies that behavior the estimator $\hat{\boldsymbol{\theta}}$ is expected to be similar to that of the MLE in the limit of $\alpha \to 1$ and $\alpha' \to 0$.

5.2 Relationship with α-Divergence

The α-divergence between two positive measures $f, g \in \mathcal{M}$ is defined as

$$D_\alpha(f,g) = \frac{1}{\alpha(1-\alpha)} \left\langle \alpha f + (1-\alpha)g - f^\alpha g^{1-\alpha} \right\rangle .$$

Note that $D_\alpha(f,g) \geq 0$ and 0 if and only if $f = g$, and the α-divergence reduces to $\mathrm{KL}(f,g)$ and $\mathrm{KL}(g,f)$ in the limit of $\alpha \to 1$ and 0, respectively. See Amari and Nagaoka (2000) for details of the α-divergences.

Remark 10 *The estimator defined by minimizing α-divergence $D_\alpha(\tilde{p}, \bar{q}_{\boldsymbol{\theta}})$ between the empirical distribution $\widetilde{p}$ and normalized model $\bar{q}$ satisfies*

$$\frac{\partial D_\alpha(\tilde{p}, \bar{q}_{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \propto \left\langle \tilde{p}^\alpha q_{\boldsymbol{\theta}}^{1-\alpha} \left(\psi'_{\boldsymbol{\theta}} - \left\langle \bar{q}_{\boldsymbol{\theta}} \psi'_{\boldsymbol{\theta}} \right\rangle \right) \right\rangle = 0,$$

which requires the computation of the normalizing constant. The same holds when unnormalized models are used. Hence, naive application of the α-divergences does not resolve the computational issue.

Here, we assume that $\alpha, \alpha' \neq 0, 1$ and consider a trick to cancel out the term $\langle g \rangle$ by mixing two α-divergences as follows.

$$\begin{aligned} D_{\alpha,\alpha'}(f,g) =& D_\alpha(f,g) + \left(\frac{-\alpha'}{\alpha} \right) D_{\alpha'}(f,g) \\ =& \left\langle \left(\frac{1}{1-\alpha} - \frac{\alpha'}{\alpha(1-\alpha')} \right) f - \frac{1}{\alpha(1-\alpha)} f^\alpha g^{1-\alpha} + \frac{1}{\alpha(1-\alpha')} f^{\alpha'} g^{1-\alpha'} \right\rangle . \end{aligned}$$

Note that $D_{\alpha,\alpha'}(f,g) \geq 0$ is divergence when $\alpha\alpha' < 0$ holds, i.e., $D_{\alpha,\alpha'}(f,g) \geq 0$ and $D_{\alpha,\alpha'}(f,g) = 0$ if and only if $f = g$. Without loss of generality, we assume that $\alpha' < 0 < \alpha \neq 1$ holds for the parameter of $D_{\alpha,\alpha'}$.

Firstly, let us consider the estimator obtained by the minimizer of $D_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}})$,

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \operatorname*{argmin}_{\boldsymbol{\theta}} D_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}}) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \in \mathcal{Z}} \frac{1}{\alpha} \left\{ \frac{1}{1-\alpha'} \left(\frac{n_{\boldsymbol{x}}}{n} \right)^{\alpha'} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha'} - \frac{1}{1-\alpha} \left(\frac{n_{\boldsymbol{x}}}{n} \right)^{\alpha} q_{\boldsymbol{\theta}}(\boldsymbol{x})^{1-\alpha} \right\} . \end{aligned} \tag{16}$$

Note that the summation in (16) includes only $\mathcal{Z}(\leq n)$ terms. Let $\bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ be the underlying distribution and $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ be the unnormalized model (2). Then the above estimator is not Fisher consistent, i.e.,

$$\left. \frac{\partial D_{\alpha,\alpha'}(\bar{q}_{\boldsymbol{\theta}_0}, q_{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \propto \left\langle \bar{q}_{\boldsymbol{\theta}_0}^{\alpha'} q_{\boldsymbol{\theta}_0}^{1-\alpha'} \psi'_{\boldsymbol{\theta}_0} - \bar{q}_{\boldsymbol{\theta}_0}^{\alpha} q_{\boldsymbol{\theta}_0}^{1-\alpha} \psi'_{\boldsymbol{\theta}_0} \right\rangle = \left(\langle q_{\boldsymbol{\theta}_0} \rangle^{-\alpha'} - \langle q_{\boldsymbol{\theta}_0} \rangle^{-\alpha} \right) \left\langle q_{\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \right\rangle \neq 0.$$

Hence, the estimator associated with $D_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}})$ does not have suitable properties such as (asymptotic) unbiasedness and consistency, while computational cost is drastically reduced. Intuitively, this is because the divergence $D_{\alpha,\alpha'}$ satisfies the coincidence axiom, i.e., $D_{\alpha,\alpha'}(f,g) = 0$ leads to $f = g$.

The statistical consistency is recovered by employing a homogeneous divergence derived from $D_{\alpha,\alpha'}$. Let us consider the estimator defined by

$$(\hat{\boldsymbol{\theta}}, \hat{z}) = \operatorname*{argmin}_{\boldsymbol{\theta},\, z>0} D_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}}/z), \tag{17}$$

where $\hat{z}$ is an estimator of the normalization term $Z_{\boldsymbol{\theta}} = \langle q_{\boldsymbol{\theta}} \rangle$.

Proposition 11 *Let $q_{\boldsymbol{\theta}}(\boldsymbol{x})$ be the unnormalized model* (2)*. Then, the minimum solution of*

$$\min_{\boldsymbol{\theta}} \min_{z>0} D_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}}/z)$$

with respect to $\boldsymbol{\theta}$ is equal to that of

$$\operatorname{sgn}(\alpha-1) S_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}}),$$

where $S_{\alpha,\alpha'}$ is formally defined for (α, α') such that $\alpha' < 0 < \alpha \neq 1$.

Proof For a given $\boldsymbol{\theta}$, we observe that

$$\hat{z} = \operatorname*{argmin}_{z>0} D_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}}/z) = \left(\frac{\left\langle \tilde{p}^{\alpha'} q_{\boldsymbol{\theta}}^{1-\alpha'} \right\rangle}{\left\langle \tilde{p}^{\alpha} q_{\boldsymbol{\theta}}^{1-\alpha} \right\rangle} \right)^{\frac{1}{\alpha-\alpha'}}. \tag{18}$$

Note that computation of (18) requires only sample order $\mathcal{O}(n)$ calculation. Then, we have

$$\begin{aligned} \min_{z>0} D_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}}/z) &= \frac{\alpha-\alpha'}{\alpha(1-\alpha')(\alpha-1)} \left(\left\langle \tilde{p}^{\alpha} q_{\boldsymbol{\theta}}^{1-\alpha} \right\rangle^{\frac{1-\alpha'}{\alpha-\alpha'}} \left\langle \tilde{p}^{\alpha'} q_{\boldsymbol{\theta}}^{1-\alpha'} \right\rangle^{\frac{\alpha-1}{\alpha-\alpha'}} - 1 \right) \\ &= \frac{\alpha-\alpha'}{\alpha(1-\alpha')(\alpha-1)} \left(e^{S_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}})} - 1 \right). \end{aligned} \tag{19}$$

for $\alpha' < 0 < \alpha \neq 1$. ■

If $\alpha' < 0$ and $1 < \alpha$ hold, the estimator using a homogeneous variant of $D_{\alpha,\alpha'}$ is equivalent to the estimator associated with the localized PS-divergence $S_{\alpha,\alpha'}$. For α and α' such that $\alpha' < 0 < \alpha < 1$, one can verify that $\operatorname{sgn}(\alpha-1) S_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}})$ is not convex in the parameter $\boldsymbol{\theta}$ of the unnormalized exponential model $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\{\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x})\}$, though $\operatorname{sgn}(\alpha-1) S_{\alpha,\alpha'}$ is still a homogeneous divergence. In any case, the mixture of α-divergences is expressed by $S_{\alpha,\alpha'}$.

From a viewpoint of the information geometry (Amari and Nagaoka, 2000), α-divergences induce the Fisher metric that is an information geometrical structure on statistical manifolds. In other word, the Hessian matrix of the α-divergences is nothing but the Fisher information matrix. This implies that the estimation based on the (mixture of) α-divergences is Fisher efficient and is an intuitive explanation of Theorem 7. The localized PS divergences, $S_{\alpha,\alpha',\gamma}$ and $S_{\alpha,\alpha'}$, and its extension $S_{\boldsymbol{\alpha},\boldsymbol{\delta}}$ in (12) can be interpreted as an extension of the α-divergences while keeping the Fisher efficiency.

We can estimate the normalization constant $Z_{\hat{\boldsymbol{\theta}}}$ by the optimal solution $\hat{z}$ of (18).

Theorem 12 *The asymptotic distribution of the estimator* (17) *is given as a degenerated multivariate normal distribution,*

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0,\, \log \hat{z} - \log Z_{\boldsymbol{\theta}_0}) \;\sim\; N(\mathbf{0}, V),$$

where V is written as

$$V = \begin{pmatrix} I_{\boldsymbol{\theta}_0}^{-1} & I_{\boldsymbol{\theta}_0}^{-1}(\log Z_{\boldsymbol{\theta}_0})' \\ (\log Z_{\boldsymbol{\theta}_0})'^T I_{\boldsymbol{\theta}_0}^{-1} & (\log Z_{\boldsymbol{\theta}_0})'^T I_{\boldsymbol{\theta}_0}^{-1}(\log Z_{\boldsymbol{\theta}_0})' \end{pmatrix}.$$

Proof is shown in Appendix E.

6. Experiments

We especially focus on a setting of $\bar{\alpha} = 1$, i.e., convexity of the risk function with the unnormalized model $\exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}))$ holds (Theorem 4), and examined the performance of the proposed estimator.

6.1 Fisher Efficiency

We investigate the Fisher efficiency of the estimator using the localized PS-divergence. In numerical experiments, training samples were generated from the Poisson distribution having the probability function

$$p_\theta(x) = \frac{e^{x\theta - e^\theta}}{x!}, \quad \theta \in \mathbb{R}$$

for $x = 0, 1, 2, \ldots$, where θ is the natural parameter. The usual parameter λ of the Poisson distribution is written as $\lambda = e^\theta$ that is equal to the expectation of x. Given the i.i.d. data $x_1, \ldots, x_n$, the MLE the parameter θ is given as $\log \bar{x}$ using the empirical mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

We used the unnormalized model defined as

$$q_\theta(x) = \frac{e^{x\theta}}{x!}.$$

Theorem 4 ensures that the localized PS-divergence with $\bar{\alpha} = 1$ from the empirical distribution $\tilde{p}$ to the above unnormalized model, i.e., $S_{\alpha,\alpha'}(\tilde{p}, q_\theta)$, is convex in θ. In numerical experiments, the parameters of $S_{\alpha,\alpha'}$ was set to $\alpha = 1.1, \alpha' = 0.1$. In addition, the estimator using a pair scoring rule (Dawid et al., 2012),

$$\hat{\theta}_\kappa = \frac{\sum_{x\geq 0} \tilde{p}(x+1)(x+1)^{1+\kappa}}{\sum_{x\geq 0} \tilde{p}(x)(x+1)^{\kappa}},$$

was compared to the MLE and proposed method. Note that the above estimator with $\kappa = 0$ is nothing but the MLE. The estimator with $\kappa = 1$ was used.

Numerical results are presented in Figure 1. The horizontal axis is the number of sample size n and the vertical axis is the averaged mean square errors of the estimated parameter from the true parameter θ_0 multiplied by n, i.e., $n \cdot \mathbb{E}[(\hat{\theta} - \theta_0)^2]$. Numerically, the

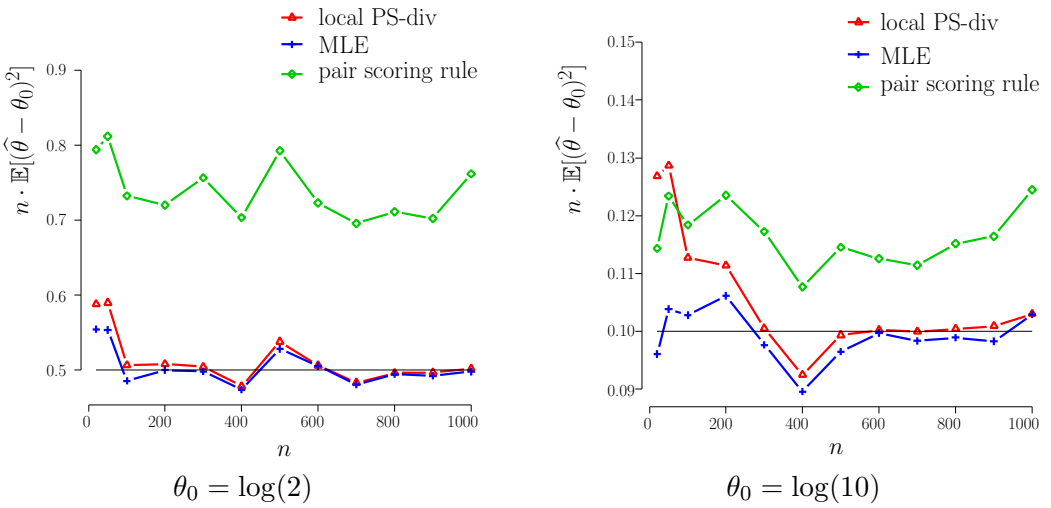


Figure 1: Mean square errors of the estimator using localized PS-divergence, MLE, and pair scoring rule are shown. Horizontal axis is the sample size and vertical axis is mean square errors multiplied by the sample size. Left panel: $\theta_0 = \log(2)$. Right panel: $\theta_0 = \log(10)$.

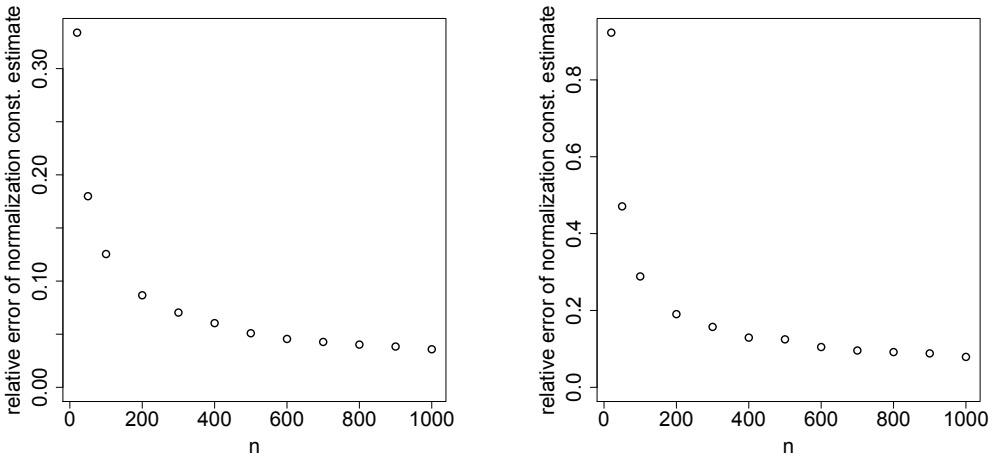
averaged mean square errors over 1000 repetitions were computed. The horizontal solid line denotes the Cramér-Rao bound. All the estimators are consistent, since the averaged mean square errors seems to be of the order $O(1/n)$. Numerically, we showed that the localized PS-divergence produces a Fisher efficient estimator, while the pair scoring rule does not achieve the Cramér-Rao bound.

When the variance of the data, i.e., e^{θ_0}, was large, the MSE of the proposed estimator for the small sample size became large. This result indicates that the regularization will be needed when a large-scale unnormalized model is used.

In addition to the parameter estimation, the normalization constant, $Z_\theta = e^{e^\theta}$, was also estimated using the equation (18). The results are presented in Fig. 2, which shows the averaged relative error of the estimator of the normalization constant, $|\hat{z} - Z_{\theta_0}|/Z_{\theta_0}$, over 1000 repetitions. Numerically, the convergence speed of the estimator (18) was approximately of the order $O(1/\sqrt{n})$.

6.2 Fully Visible Boltzmann Machine

In this subsection, we compared the proposed estimator with parameter settings $(\alpha, \alpha') = (1.01, 0.01), (1.01, -0.01), (2, -1)$, with the MLE and the ratio matching method (Hyvärinen, 2007). Note that the ratio matching method also does not require calculation of the normalization constant. In this experiment, we do not compare the proposed estimator with the pseudo-likelihood method and the contrastive divergence which also do not require the calculation of the normalization constant, because in (Hyvärinen, 2007), the ratio matching method has been numerically compared with the pseudo-likelihood method and the con-



$\theta_0 = \log(2)$ $\theta_0 = \log(10)$

Figure 2: Averaged relative errors of the estimator (18) for the normalization constant are shown. Horizontal axis is the sample size. Left panel: $\theta_0 = \log(2)$. Right panel: $\theta_0 = \log(10)$.

trastive divergence using the fully visible Boltzmann machine, and its result showed that all methods have almost equal performance. In addition, some theoretical properties of the pseudo-likelihood (and its extension, the composite likelihood) have been investigated in (Mardia et al., 2009; Kanamori, 2016), showing that the pseudo-likelihood is not Fisher efficient in general.

All methods were optimized with the *optim* function in R language (R Core Team, 2015). The dimension d of input was set to 10 and the synthetic dataset was randomly generated from the second order Boltzmann machine (Example 3) with a parameter $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, I/d)$. We repeated comparison 50 times and observed averaged performance. Figure 3 (a) shows median of the root mean square errors (RMSEs) between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ of each method over 50 trials, against the number n of examples. We observe that the proposed estimator is comparable with the MLE as predicted by the Theorem 7, and the ratio matching is also comparable to the MLE under the setting. Figure 3 (b) shows a number of observed patterns in a dataset consists of n examples. Figure 3 (c) shows median of computational time of each method against n. The computational time of the MLE does not vary against n because the computational cost is dominated by the calculation of the normalization constant. Both the proposed estimator and the ratio matching method are significantly faster than the MLE.

We investigated performance of methods under an another situation, in which the true parameter distributes as $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, I)$. Figure 4 (a) shows median of the root mean square errors (RMSEs) between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$ of each method over 50 trials, against the number n of examples. We observe that the proposed estimator works well, but the MLE outperforms the proposed method contrary to the prediction of Theorem 7. This is because observed

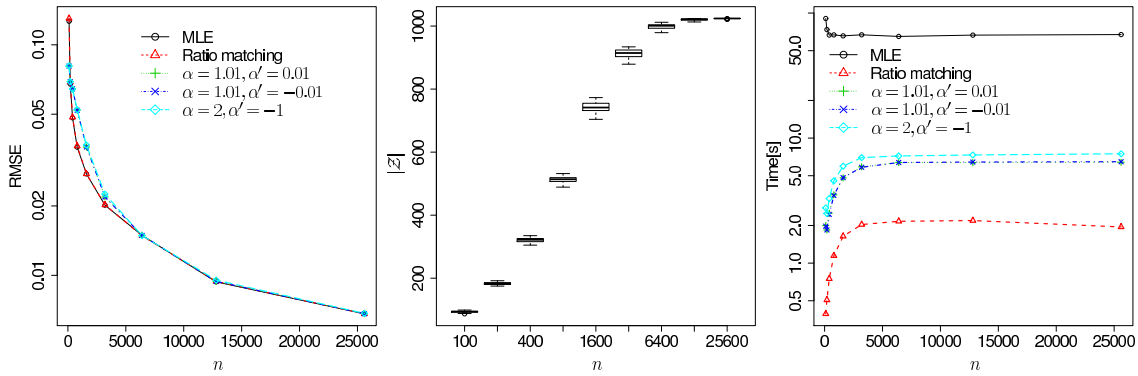


Figure 3: Results for datasets generated with a parameter $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, I/d)$. (a) Median of RMSEs of each method against n, in log scale. (b) Box-whisker plot of number $|\mathcal{Z}|$ of unique patterns in the dataset $\mathcal{D}$ against n. (c) Median of computational time of each method against n, in log scale.

patterns were only a small portion of all possible patterns, as shown in Figure 4 (b). Even in such a case, the MLE can take all possible patterns ($2^{10} = 1024$) into account through the normalization term $Z_{\boldsymbol{\theta}}$ because the Taylor expansion of $\log Z_{\boldsymbol{\theta}}$ around $\boldsymbol{\theta} = \mathbf{0}$ which is approximated as

$$\log Z_{\boldsymbol{\theta}} \simeq d \log 2 + \frac{1}{2}||\boldsymbol{\theta}||^2$$

term behaves like a regularizer. On the other hand, the proposed method genuinely uses only the observed examples, and focuses on the restricted domain $\mathcal{Z}$ rather than the original domain $\mathcal{X}$, in which the asymptotic analysis would not be relevant in this case. Figure 4 (c) shows median of computational time of each method against n. Both the proposed estimator and the ratio matching method are significantly faster than the MLE. While the ratio matching method is faster than the proposed estimator, the RMSE of the proposed estimator is less than that of the ratio matching.

To overcome the degrade of performance of the proposed estimator caused by lack of example patterns, we consider a regularized version of the proposed estimator as,

$$\operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ S_{\alpha,\alpha'}(\tilde{p}, q_{\boldsymbol{\theta}}) + \frac{\lambda}{2n}||\boldsymbol{\theta}||^2 \right\}. \tag{20}$$

Note that we can employ the l_1 regularizer to obtain a sparse estimator, rather than the l_2 regularizer. Ravikumar et al. (2010) theoretically investigated conditions for correctly selecting edges of the Boltzmann machine using the l_1-regularized logistic regression.

We investigated performance of the regularized estimator with the same dataset ($\boldsymbol{\theta}^* \sim \mathcal{N}(0, I)$). Figure 5 shows median of RMSEs and computational time of the MLE, the ratio matching, and the regularized estimator ($\alpha = 1.01, \alpha' = 0.01$, $\lambda = 0, 10^{-2}, 10^{-4}$). The Figure shows that performance of the proposed estimator is improved by the regularization term $||\boldsymbol{\theta}||^2$, which can be interpreted as an approximation of the normalization constant $\log Z_{\boldsymbol{\theta}}$. Note that the computational time of the regularized estimator is drastically reduced

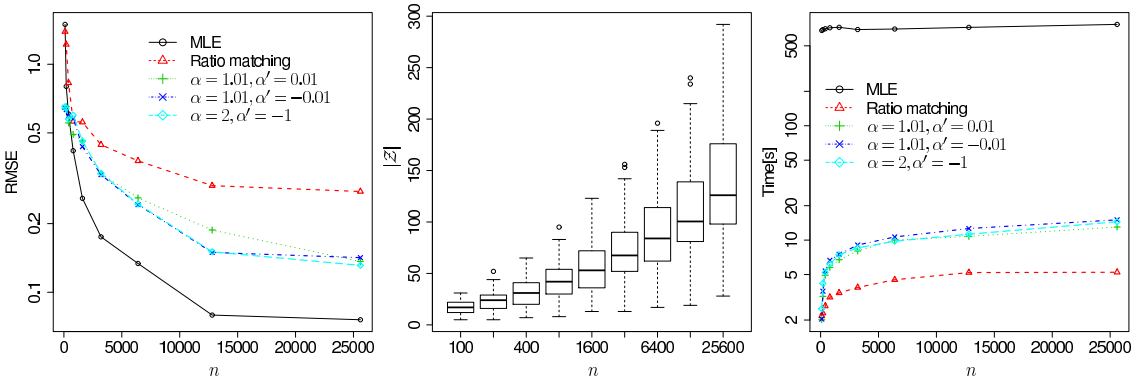


Figure 4: Results for datasets generated with a parameter $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, I)$. (a) Median of RMSEs of each method against n, in log scale. (b) Box-whisker plot of number $|\mathcal{Z}|$ of unique patterns in the dataset $\mathcal{D}$ against n. (c) Median of computational time of each method against n, in log scale.

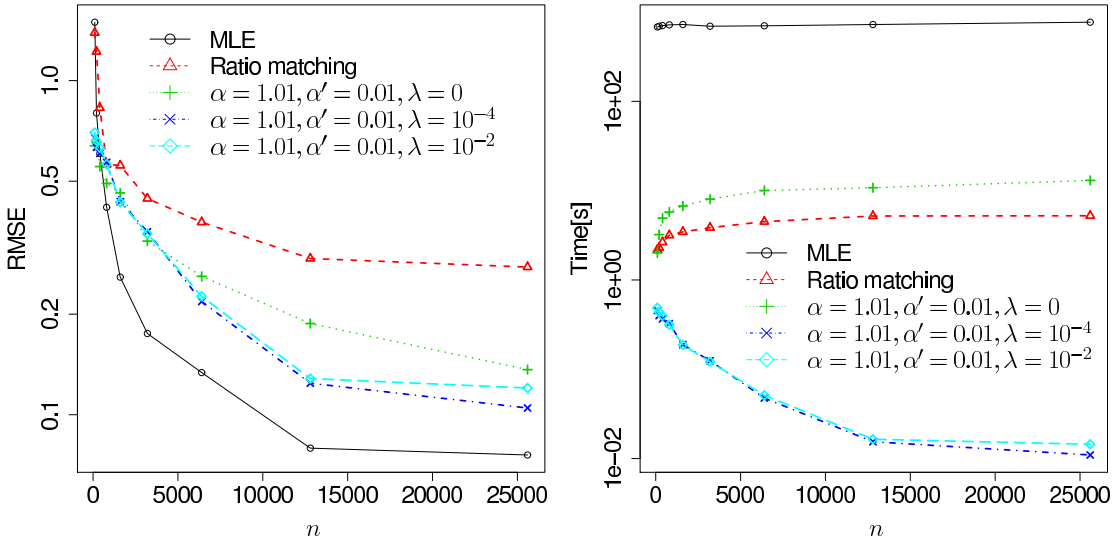


Figure 5: (a) Median of RMSE of the MLE, the ratio matching and the regularized estimator with $\alpha = 1.01, \alpha' = 0.01$, and $\lambda = 0, 10^{-2}, 10^{-4}$ against n, in log scale. (b) Median of computational time of each method against n, in log scale.

compared with that of estimator without regularization. This is because the cost function (20) becomes similar to a quadratic function and the condition number associated with the cost function is improved by the regularization, which influences required number of steps of the quasi-Newton method.

Also we investigated performance of the estimator (18) for the normalization constant $Z_{\boldsymbol{\theta}}$ and results for above two situations are shown in Figure 6. In the Figure 6, the averaged relative errors $|\hat{z} - Z_{\boldsymbol{\theta}_0}| / Z_{\boldsymbol{\theta}_0}$ of the proposed estimator $((\alpha, \alpha') = (1.01, 0.01), (1.01, -0.01), (2, -1))$ are shown. We observe that the proposed estimator appropriately works.

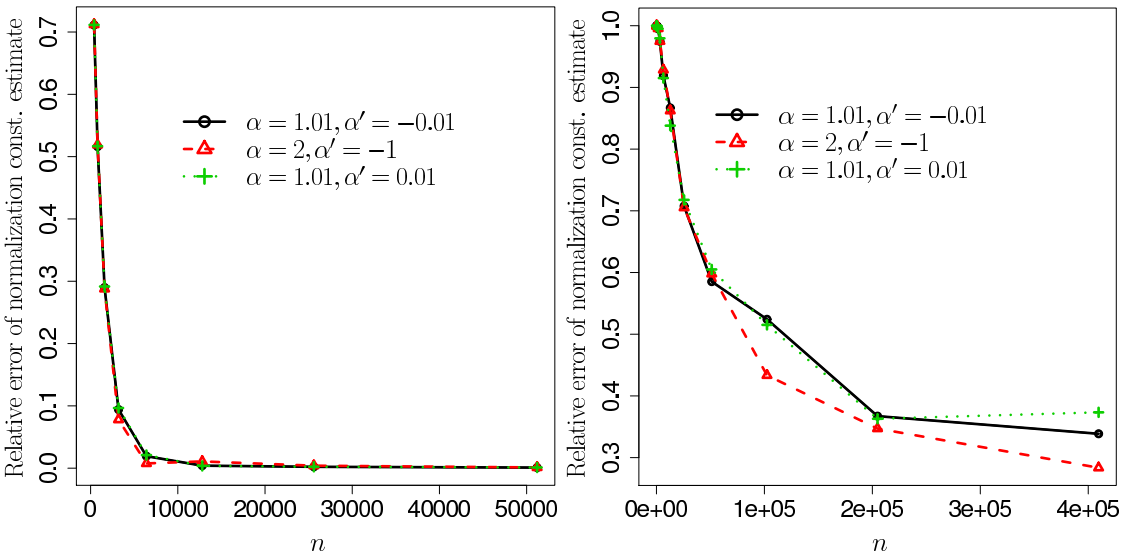


Figure 6: Median of relative errors of normalization constant estimates for two experiment scenarios: (a) $\boldsymbol{\theta}^* \sim \mathcal{N}(0, I/d)$. (b) $\boldsymbol{\theta}^* \sim \mathcal{N}(0, I)$.

6.3 Sample Complexity

We numerically investigated the sample complexity of the proposed estimator ($\alpha_1 = 1.01, \alpha' = 0.01$, without the regularization term), i.e., a number $N(\epsilon, \delta, d)$ of examples to attain

$$\Pr(\mathrm{KL}(\bar{q}_{\boldsymbol{\theta}_0}, \bar{q}_{\hat{\boldsymbol{\theta}}}) \geq \epsilon) < \delta$$

where ϵ and δ are arbitrary positive constants. For each dimension $d = 2, 3, \ldots, 21$, we generated a dataset containing $n = 50 \times 2^k (k = 1, \ldots, 9)$ examples from the fully visible Boltzmann machine and calculated the KL divergences $\mathrm{KL}(\bar{q}_{\boldsymbol{\theta}_0}, \bar{q}_{\hat{\boldsymbol{\theta}}})$. For two kinds of parameter settings, i.e., $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, I/d)$ and $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, I)$, we repeated these procedures 50 times and observed 50 values of KL divergence. Each panel in Figure 7 shows median of values of the KL divergence ($\delta = 0.5$) for each setting of the parameter, respectively. Both panels show that the proposed estimator requires approximately $2 \sim 3$ times as many examples examples against an increase of the dimensionality d at the same level ϵ of estimation error $\mathrm{KL}(\bar{q}_{\boldsymbol{\theta}_0}, \bar{q}_{\hat{\boldsymbol{\theta}}})$.

6.4 Boltzmann Machine with Hidden Variables

In this subsection, we applied the proposed estimator for the Boltzmann machine with hidden variables whose associated function is written as (3). The proposed estimator with parameter settings $(\alpha, \alpha') = (1.01, 0.01), (1.01, -0.01), (2, -1)$ was compared with the MLE. The dimension d_1 of observed variables was fixed to 10 and d_2 of hidden variables was set to 2, and the parameter $\boldsymbol{\theta}^*$ was generated as $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, I)$ including parameters corresponding to hidden variables. Note that the Boltzmann machine with hidden variables is not identifiable and different values of the parameter do not necessarily generate different probability distributions, implying that estimators are influenced by local minimums. Then we

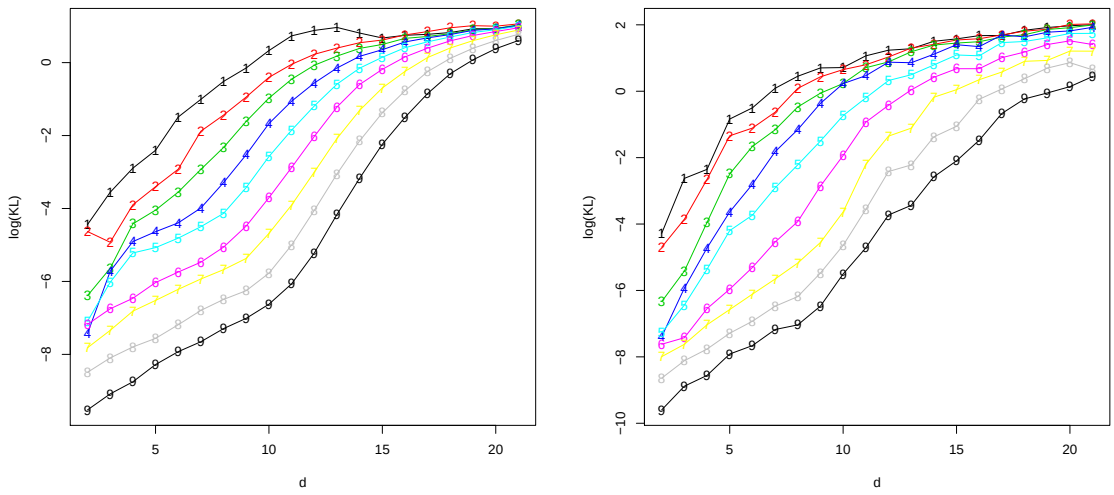


Figure 7: Median of KL divergences $\mathrm{KL}(\bar{q}_{\boldsymbol{\theta}_0}, \bar{q}_{\hat{\boldsymbol{\theta}}})$. Horizontal axis is dimension d, and vertical axis is median of KL divergences, in log-scale. Number k ($k = 1, \ldots, 9$) on a line corresponds to number $n = 50 \times 2^k$ of examples. (a) $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, I/d)$. (b) $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, I)$,

measured performance of each estimator by the averaged log-likelihood $\frac{1}{n}\sum_{i=1}^{n} \log \bar{q}_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}_i)$ rather than the RMSE of parameters. An initial value of the parameter was set by $\mathcal{N}(\mathbf{0}, I)$ and commonly used by all methods. We repeated the comparison 50 times and observed the averaged performance. Figure 8 (a) shows median of averaged log-likelihoods of each method for the training dataset over 50 trials, against the number n of example. We observe that the proposed estimator is comparable with the MLE when the number n of examples becomes large. Note that the averaged log-likelihood of MLE once decreases when n is small, and this is due to overfitting of the model. Figure 8 (b) shows median of averaged log-likelihoods of each method for test dataset consists of 10000 examples, over 50 trials. Figure 8 (c) shows median of computational time of each method against n, and we observe that the proposed estimator is significantly faster than the MLE.

In addition, we investigated performance of the regularized version (20) ($\alpha = 1.01, \alpha' = 0.01$ and $\lambda = 0, 10^{-2}, 10^{-4}$) and compared with the MLE. Figures 9 (a) and (b) show medians of averaged log-likelihoods of the MLE and the regularized estimator over 50 trials, for the training dataset and the test dataset, respectively. Figures imply that an appropriate regularization can improve performance of the estimator and also as the experiment in previous subsection, computational time of the regularized estimator is drastically reduced, in comparison with that of the estimator without regularization.

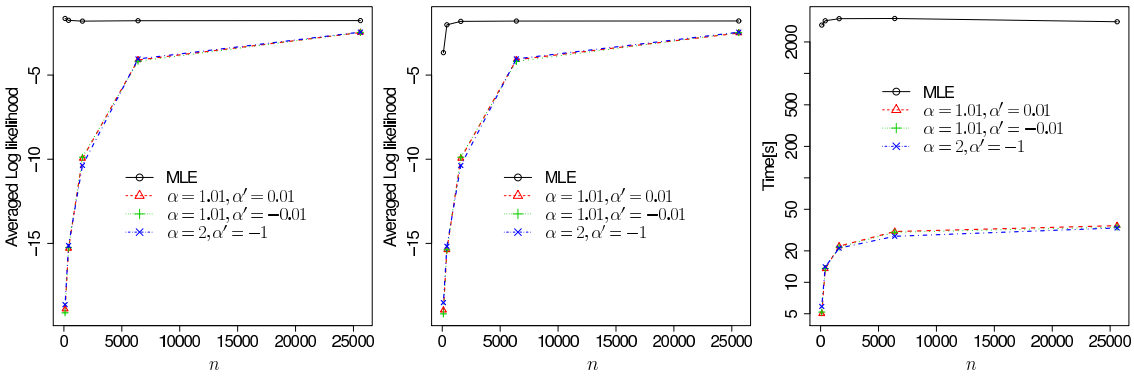


Figure 8: (a) Median of averaged log-likelihoods of each method against n. (b) Median of averaged log-likelihoods of each method calculated for test dataset against n. (c) Median of computational time of each method against n, in log scale.

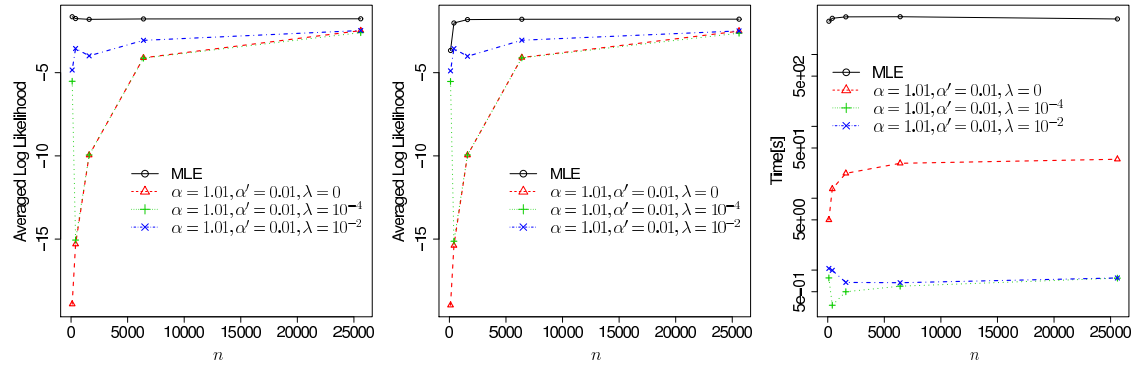


Figure 9: (a) Median of averaged log-likelihoods of the MLE and regularized estimator with $\lambda = 0, 10^{-2}, 10^{-4}$ against n. (b) Median of averaged log-likelihoods of each method calculated for test dataset against n. (c) Median of computational time of each method against n, in log scale.

7. Conclusions

We proposed a novel estimator for probabilistic models on discrete space, for which the normalization constant is infeasible to calculate. The proposed estimator is based on unnormalized models and an empirically localized PS-divergence having the homogeneous property and can be constructed without calculation of the normalization constant. We showed that homogeneous divergences with empirical localization allow the computation of the normalization constant to be avoided because of a weak coincidence axiom. The idea of empirical localization permits ignoring an unobserved domain on sample space, which can drastically reduce computational cost. We investigated statistical properties of the proposed estimator and revealed that the proposed estimator is asymptotically efficient and that its asymptotic distribution is equal to that of the maximum likelihood estimator (MLE). In addition, we showed a relationship between the empirically localized PS-divergence and a mixture of α-divergences. The Hessian matrix of the α-divergence is known to be equal

to the Fisher information matrix, which implies the Fisher efficiency of the proposed estimator. A difference between two divergences is that the mixture of α-divergence does not have homogeneous property. By utilizing this difference, we proposed an estimator for the normalization constant that requires only sample order $\mathcal{O}(n)$ calculation and is asymptotically consistent. We investigated the performance of the proposed estimator with various kinds of models on discrete sample space and showed that the proposed estimator performs comparably to the MLE, while required computational cost is drastically reduced.

A possible future direction for this work is application of the proposed framework for models on continuous space. While employment of the framework for continuous space is an interesting challenge, the empirical localization technique is difficult to apply to models on continuous space because of the power of the empirical distribution described by the delta function. Another direction is making the proposed estimator more robust. The PS-divergence used in this paper is well known to be robust against outlier noise (Kanamori and Fujisawa, 2015). The robustness of the PS-divergence can be controlled by the tuning parameter γ, which was fixed to a value ensuring convexity of the risk function in this paper. The robustification of the proposed estimator and its theoretical justification are important issues for data analysis application.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 25730018, 16K00051, 16K00044, 15H01678, and 15H03636 from MEXT, Japan.

Appendix A. Proof of Theorem 4

Some calculation yields

$$\frac{\partial^2 \log \left\langle p^\alpha q_{\boldsymbol{\theta}}^{1-\alpha} \right\rangle}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = (1-\alpha)^2 V_{r_{\alpha,\boldsymbol{\theta}}}[\boldsymbol{\phi}],$$

where $V_{r_{\alpha,\boldsymbol{\theta}}}[\boldsymbol{\phi}]$ is the covariance matrix of $\boldsymbol{\phi}(\boldsymbol{x})$ under the probability $r_{\alpha,\boldsymbol{\theta}}(\boldsymbol{x})$. Thus, the Hessian matrix of $S_{\alpha,\alpha',\gamma}(p, q_{\boldsymbol{\theta}})$ is written as

$$\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} S_{\alpha,\alpha',\gamma}(p, q_{\boldsymbol{\theta}}) = \frac{(1-\alpha)^2}{1+\gamma} V_{r_{\alpha,\boldsymbol{\theta}}}[\boldsymbol{\phi}] + \frac{\gamma(1-\alpha')^2}{1+\gamma} V_{r_{\alpha',\boldsymbol{\theta}}}[\boldsymbol{\phi}] - (1-\bar{\alpha})^2 V_{r_{\bar{\alpha},\boldsymbol{\theta}}}[\boldsymbol{\phi}].$$

The Hessian matrix is positive semidefinite if $\bar{\alpha} = 1$.

To prove the second part, we prove that there exists a distribution p, a model $q_{\boldsymbol{\theta}}$ and parameters α, α', γ such that the Hessian is not positive semidefinite, for a given $\bar{\alpha} \neq 1$. Suppose that $\mathcal{X} = \{+1, -1\}^d$ and the function $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_d(\boldsymbol{x})) \in \mathbb{R}^d$ is defined by $\phi_k(\boldsymbol{x}) = x_k$, $k = 1, \ldots, d$ for $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X}$. Let p be the uniform distribution on $\mathcal{X}$. The covariance matrix of $\boldsymbol{\phi}$ is the diagonal matrix given by

$$V_{r_{\alpha,\boldsymbol{\theta}}}[\boldsymbol{\phi}] = 4 \cdot \mathrm{diag}\left(\frac{1}{(e^{(1-\alpha)\theta_1} + e^{-(1-\alpha)\theta_1})^2}, \ldots, \frac{1}{(e^{(1-\alpha)\theta_d} + e^{-(1-\alpha)\theta_d})^2} \right).$$

Let δ be $\delta = 1/(1+\gamma)$, then $\delta \in (0,1)$ holds for $\gamma > 0$. We define

$$f(z;\theta) = \frac{(1-z)^2}{(e^{(1-z)\theta} + e^{-(1-z)\theta})^2}, \quad z,\, \theta \in \mathbb{R}.$$

Then, the i-th diagonal element of the Hessian matrix is expressed by

$$\Delta = \delta \cdot f(\alpha;\theta_i) + (1-\delta)\cdot f(\alpha';\theta_i) - f(\delta\alpha + (1-\delta)\alpha';\theta_i)$$

up to a positive constant. Our task is to find the parameter α, α', δ such that $\bar{\alpha} = \delta\alpha + (1-\delta)\alpha'$ and $\Delta < 0$ hold. The function f satisfies the following properties.

(a) $f(z;\theta) \geq 0$ and $f(z;\theta) = 0 \Leftrightarrow z = 1$.

(b) $f(1+\varepsilon;\theta) = f(1-\varepsilon;\theta) = f(1+\varepsilon;-\theta)$ for $\varepsilon \geq 0, \theta \in \mathbb{R}$.

(c) $\lim_{z\to\pm\infty} f(z;\theta) = 0$ holds for $\theta \neq 0$.

Let θ be a fixed non-zero real number. Since $\bar{\alpha} \neq 1$, $f(\bar{\alpha};\theta) > 0$ holds. Due to the properties (b) and (c), any sufficiently large $\varepsilon > 0$ satisfies $f(1-\varepsilon;\theta) = f(1+\varepsilon;\theta) < f(\bar{\alpha};\theta)$. Define $\alpha = 1+\varepsilon$ and $\alpha' = 1-\varepsilon$. By choosing $\delta \in (0,1)$ such that $\bar{\alpha} = \delta\alpha + (1-\delta)\alpha'$, we have $\Delta < 0$.

Appendix B. Proof of Proposition 5

Suppose that $\bar{\alpha} > 1$. Due to (a), (c) in Appendix A and the continuity of $f(z,\theta)$ at $z = 1$, there exists α and α' satisfying $1 < \alpha' < \bar{\alpha} < \alpha$ such that both $f(\alpha';\theta)$ and $f(\alpha;\theta)$ are less than $f(\bar{\alpha};\theta)$, where θ is a non-zero constant. Then, $\Delta < 0$ holds for $\delta \in (0,1)$ such that $\bar{\alpha} = \delta\alpha + (1-\delta)\alpha'$. We prove the case of $0 < \bar{\alpha} < 1$. For a sufficiently large θ, we have

$$\frac{f(0;\theta)}{f(\bar{\alpha};\theta)} = O\left(\frac{e^{-2\bar{\alpha}\theta}}{(1-\bar{\alpha})^2}\right) \to 0 \ (\theta \to \infty).$$

Hence, the continuity of f ensures that there exist a sufficiently large θ and a small positive α' such that $0 < \alpha' < \bar{\alpha}$ and $f(\alpha';\theta) < f(\bar{\alpha};\theta)$ hold. The property (c) in the Appendix A ensures that there exists a sufficiently large $\alpha > 1$ satisfying $f(\alpha;\theta) < f(\bar{\alpha};\theta)$. Again, $\Delta < 0$ holds for δ such that $\bar{\alpha} = \delta\alpha + (1-\delta)\alpha'$.

Appendix C. Generalized Hölder's inequality

This section includes the proof and the equality condition of the generalized Hölder's inequality. The inequality (11) of $L = 2$ is nothing but the standard Hölder's inequality. The inequality (11) of $L = 3$ is proved by using the standard one as follows:

$$\left\langle f_1^{\delta_1} f_2^{\delta_2} f_3^{\delta_3} \right\rangle = \left\langle (f_1^{\frac{\delta_1}{1-\delta_3}} f_2^{\frac{\delta_2}{1-\delta_3}})^{1-\delta_3} f_3^{\delta_3} \right\rangle \leq \left\langle f_1^{\frac{\delta_1}{1-\delta_3}} f_2^{\frac{\delta_2}{1-\delta_3}} \right\rangle^{1-\delta_3} \langle f_3\rangle^{\delta_3} \leq \langle f_1\rangle^{\delta_1} \langle f_2\rangle^{\delta_2} \langle f_3\rangle^{\delta_3},$$

where $\delta_1 + \delta_2 + \delta_3 = 1$ and $\delta_i > 0$ for $i = 1,2,3$ are assumed. The third inequality becomes equality if and only if f_1 and f_2 are linearly dependent, and the second inequality becomes

equality if and only if $f_1^{\frac{\delta_1}{1-\delta_3}} f_2^{\frac{\delta_2}{1-\delta_3}}$ and f_3 are linearly dependent. As a result, (11) of $L=3$ becomes an equality if and only if there exists a function $g \in \mathcal{M}$ such that all f_1, f_2, f_3 are proportional to g. In the same way, we can prove the generalized Hölder's inequality of any natural number L.

Appendix D. Proof of Theorem 7

We show the asymptotic distribution of the estimator defined from the localized PS-divergence $S_{\alpha,\alpha',\gamma}$. Along the same line, one can prove the asymptotic property of the estimator defined from the general localized PS-divergence $S_{\boldsymbol{\alpha},\boldsymbol{\delta}}$.

Let us assume that the empirical distribution is written as

$$\tilde{p}(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x}) + \epsilon \cdot s(\boldsymbol{x}),$$

where ϵ is a small positive number and $s(\boldsymbol{x})$ satisfies $\langle s \rangle = 0$. Note that $r_{\alpha,\boldsymbol{\theta}_0}(\boldsymbol{x}) = \bar{q}_{\boldsymbol{\theta}_0}(\boldsymbol{x})$. By expanding an equilibrium condition of the estimator (13) around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we obtain

$$\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\theta}} S_{\alpha,\alpha',\gamma}(\tilde{p}, q_{\boldsymbol{\theta}}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} S_{\alpha,\alpha',\gamma}(\tilde{p}, q_{\boldsymbol{\theta}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} S_{\alpha,\alpha',\gamma}(\tilde{p}, q_{\boldsymbol{\theta}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathcal{O}(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||^2) \\
&= \left\{ \frac{1-\alpha}{1+\gamma} \left\langle \tilde{r}_{\alpha,\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \right\rangle + \frac{\gamma(1-\alpha')}{1+\gamma} \left\langle \tilde{r}_{\alpha',\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \right\rangle - (1-\bar{\alpha}) \left\langle \tilde{r}_{\bar{\alpha},\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \right\rangle \right\} \\
&\quad + \left\{ \frac{(1-\alpha)^2}{1+\gamma} V_{\tilde{r}_{\alpha,\boldsymbol{\theta}_0}}[\psi'_{\boldsymbol{\theta}_0}] + \frac{\gamma(1-\alpha')^2}{1+\gamma} V_{\tilde{r}_{\alpha',\boldsymbol{\theta}_0}}[\psi'_{\boldsymbol{\theta}_0}] - (1-\bar{\alpha})^2 V_{\tilde{r}_{\bar{\alpha},\boldsymbol{\theta}_0}}[\psi'_{\boldsymbol{\theta}_0}] \right. \\
&\quad \left. + \frac{1-\alpha}{1+\gamma} \left\langle \tilde{r}_{\alpha,\boldsymbol{\theta}_0} \psi''_{\boldsymbol{\theta}_0} \right\rangle + \frac{\gamma(1-\alpha')}{1+\gamma} \left\langle \tilde{r}_{\alpha',\boldsymbol{\theta}_0} \psi''_{\boldsymbol{\theta}_0} \right\rangle - (1-\bar{\alpha}) \left\langle \tilde{r}_{\bar{\alpha},\boldsymbol{\theta}_0} \psi''_{\boldsymbol{\theta}_0} \right\rangle \right\} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \mathcal{O}(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||^2),
\end{aligned}$$

where $\psi'_{\boldsymbol{\theta}_0}$ and $\psi''_{\boldsymbol{\theta}_0}$ are the gradient vector and the Hessian matrix of $\psi_{\boldsymbol{\theta}_0}(\boldsymbol{x})$ with respect to the parameter $\boldsymbol{\theta}$. By the delta method (Van der Vaart, 1998), we observe that

$$\begin{aligned}
&\left(\left\langle \tilde{r}_{\alpha,\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \right\rangle - \left\langle r_{\alpha,\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \right\rangle \right) \\
&= \alpha\epsilon \frac{\left\langle \bar{q}_{\boldsymbol{\theta}_0}^{\alpha-1} q_{\boldsymbol{\theta}_0}^{1-\alpha} \psi'_{\boldsymbol{\theta}_0} s \right\rangle \left\langle \bar{q}_{\boldsymbol{\theta}_0}^{\alpha} q_{\boldsymbol{\theta}_0}^{1-\alpha} \right\rangle - \left\langle \bar{q}_{\boldsymbol{\theta}_0}^{\alpha} q_{\boldsymbol{\theta}_0}^{1-\alpha} \psi'_{\boldsymbol{\theta}_0} \right\rangle \left\langle \bar{q}_{\boldsymbol{\theta}_0}^{\alpha-1} q_{\boldsymbol{\theta}_0}^{1-\alpha} s \right\rangle}{\left\langle \bar{q}_{\boldsymbol{\theta}_0}^{\alpha} q_{\boldsymbol{\theta}_0}^{1-\alpha} \right\rangle^2} + O(\epsilon^2) \\
&= \alpha\epsilon \left(\left\langle \psi'_{\boldsymbol{\theta}_0} s \right\rangle - \left\langle \bar{q}_{\boldsymbol{\theta}_0} \psi'_{\boldsymbol{\theta}_0} \right\rangle \langle s \rangle \right) + O(\epsilon^2) \\
&= \alpha\epsilon \left\langle \psi'_{\boldsymbol{\theta}_0} s \right\rangle + O(\epsilon^2).
\end{aligned}$$

Then we have

$$\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\theta}}S_{\alpha,\alpha',\gamma}(\tilde{p},q_{\boldsymbol{\theta}})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - \frac{\partial}{\partial\boldsymbol{\theta}}S_{\alpha,\alpha',\gamma}(p,q_{\boldsymbol{\theta}})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\\
&= \frac{1-\alpha}{1+\gamma}\left(\left\langle\tilde{r}_{\alpha,\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle - \left\langle r_{\alpha,\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle\right) + \frac{\gamma(1-\alpha')}{1+\gamma}\left(\left\langle\tilde{r}_{\alpha',\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle - \left\langle r_{\alpha',\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle\right)\\
&\quad - (1-\bar{\alpha})\left(\left\langle\tilde{r}_{\bar{\alpha},\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle - \left\langle r_{\bar{\alpha},\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle\right)\\
&= \left\{\frac{1-\alpha}{1+\gamma}\left\langle\tilde{r}_{\alpha,\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle + \frac{\gamma(1-\alpha')}{1+\gamma}\left\langle\tilde{r}_{\alpha',\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle - (1-\bar{\alpha})\left\langle\tilde{r}_{\bar{\alpha},\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle\right\}\\
&\quad - \left\{1 - \frac{\alpha+\gamma\alpha'}{1+\gamma} - (1-\bar{\alpha})\right\}\left\langle\bar{q}_{\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle\\
&= \left(\alpha\frac{1-\alpha}{1+\gamma} + \alpha'\frac{\gamma(1-\alpha')}{1+\gamma} - (1-\bar{\alpha})\bar{\alpha}\right)\left\langle\psi'_{\boldsymbol{\theta}_0}\epsilon s\right\rangle + O(\epsilon^2)\\
&= -\frac{\gamma}{(1+\gamma)^2}(\alpha-\alpha')^2\left\langle\psi'_{\boldsymbol{\theta}_0}(\tilde{p}-\bar{q}_{\boldsymbol{\theta}_0})\right\rangle + O(\epsilon^2).
\end{aligned}$$

From the central limit theorem,

$$\sqrt{n}\left\langle\psi'_{\boldsymbol{\theta}_0}(\tilde{p}-\bar{q}_{\boldsymbol{\theta}_0})\right\rangle = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\psi'_{\boldsymbol{\theta}_0}(\boldsymbol{x}_i) - \left\langle\bar{q}_{\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle\right)$$

asymptotically follows the normal distribution with mean $\mathbf{0}$ and variance $V_{\bar{q}_{\boldsymbol{\theta}_0}}[\psi'_{\boldsymbol{\theta}_0}] = I_{\boldsymbol{\theta}_0}$. Also from the law of large numbers, we have

$$\begin{aligned}
&\frac{(1-\alpha)^2}{1+\gamma}V_{\tilde{r}_\alpha}[\psi'_{\boldsymbol{\theta}_0}] + \frac{\gamma(1-\alpha')^2}{1+\gamma}V_{\tilde{r}_{\alpha'}}[\psi'_{\boldsymbol{\theta}_0}] - (1-\bar{\alpha})^2V_{\tilde{r}_{\bar{\alpha}}}[\psi'_{\boldsymbol{\theta}_0}] \to \frac{\gamma}{(1+\gamma)^2}(\alpha-\alpha')^2I_{\boldsymbol{\theta}_0},\\
&\quad\frac{1-\alpha}{1+\gamma}\left\langle\tilde{r}_{\alpha,\boldsymbol{\theta}_0}\psi''_{\boldsymbol{\theta}_0}\right\rangle + \frac{\gamma(1-\alpha')}{1+\gamma}\left\langle\tilde{r}_{\alpha',\boldsymbol{\theta}_0}\psi''_{\boldsymbol{\theta}_0}\right\rangle - (1-\bar{\alpha})\left\langle\tilde{r}_{\bar{\alpha},\boldsymbol{\theta}_0}\psi''_{\boldsymbol{\theta}_0}\right\rangle\\
&\to \left(1 - \frac{\alpha+\gamma\alpha'}{1+\gamma} - (1-\bar{\alpha})\right)\left\langle\bar{q}_{\boldsymbol{\theta}_0}\psi''_{\boldsymbol{\theta}_0}\right\rangle = 0
\end{aligned}$$

in the limit of $n\to\infty$. Taking the probabilistic error in the law of large numbers into account, we obtain the equality

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) = I_{\boldsymbol{\theta}_0}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\psi'_{\boldsymbol{\theta}_0}(\boldsymbol{x}_i) - \left\langle\bar{q}_{\boldsymbol{\theta}_0}\psi'_{\boldsymbol{\theta}_0}\right\rangle\right)\right\} + o_p(1)$$

Consequently, the asymptotic distribution of the estimator is given as

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \sim \mathcal{N}(\mathbf{0}, I_{\boldsymbol{\theta}_0}^{-1}).$$

Appendix E. Proof of Theorem 12

Let us consider the asymptotic property of $\hat{z}$. The logarithm of the estimator $\widehat{z}$ is expressed as

$$\log\widehat{z} = \frac{1}{\alpha-\alpha'}\log\left\langle\tilde{p}^{\alpha'}q_{\hat{\boldsymbol{\theta}}}^{1-\alpha'}\right\rangle - \frac{1}{\alpha-\alpha'}\log\left\langle\tilde{p}^{\alpha}q_{\hat{\boldsymbol{\theta}}}^{1-\alpha}\right\rangle.$$

In the same way as the calculation in Appendix D, we have

$$\log \left\langle p^{\alpha} q_{\hat{\theta}}^{1-\alpha} \right\rangle = (1-\alpha) \log Z_{\boldsymbol{\theta}_0} + (1-\alpha)(\log Z_{\boldsymbol{\theta}_0})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|).$$

From the above equation, we have

$$\sqrt{n}(\log \hat{z} - \log Z_{\boldsymbol{\theta}_0}) = \sqrt{n}(\log Z_{\boldsymbol{\theta}_0})'^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(1).$$

Therefore, the asymptotic variance of of $\log(\hat{z})$ is given by

$$n \cdot V[\log(\hat{z})] \longrightarrow (\log Z_{\boldsymbol{\theta}_0})'^T I_{\boldsymbol{\theta}_0}^{-1} (\log Z_{\boldsymbol{\theta}_0})'.$$

We show the correlation between $\hat{\boldsymbol{\theta}}$ and $\log \hat{z}$. The asymptotic expansion above yields

$$n \cdot (\log \widehat{z} - (\log Z_{\boldsymbol{\theta}_0}))(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0))(\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0))^T (\log Z_{\boldsymbol{\theta}_0})' + o_p(1)$$

and we have

$$n \cdot \mathbb{E}[(\log \hat{z} - \log Z_{\boldsymbol{\theta}_0})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)] \longrightarrow I_{\boldsymbol{\theta}_0}^{-1} (\log Z_{\boldsymbol{\theta}_0})'.$$

Therefore, the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0, \log \hat{z} - \log Z_{\boldsymbol{\theta}_0})$ is given as the multivariate normal distribution with mean zero and the following asymptotic variance

$$\begin{aligned} n \cdot V[(\hat{\boldsymbol{\theta}}, \log \hat{z})] \;&\longrightarrow\; \begin{pmatrix} I_{\boldsymbol{\theta}_0}^{-1} & I_{\boldsymbol{\theta}_0}^{-1} (\log Z_{\boldsymbol{\theta}_0})' \\ (\log Z_{\boldsymbol{\theta}_0})'^T I_{\boldsymbol{\theta}_0}^{-1} & (\log Z_{\boldsymbol{\theta}_0})'^T I_{\boldsymbol{\theta}_0}^{-1} (\log Z_{\boldsymbol{\theta}_0})' \end{pmatrix} \\ &= \begin{pmatrix} E \\ (\log Z_{\boldsymbol{\theta}_0})'^T \end{pmatrix} I_{\boldsymbol{\theta}_0}^{-1} \begin{pmatrix} E & (\log Z_{\boldsymbol{\theta}_0})' \end{pmatrix}, \end{aligned}$$

where E is the identity matrix. Note that the distribution of the estimator $(\hat{\boldsymbol{\theta}}, \hat{z})$ is asymptotically degenerated.

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, 2000.

Shun-ichi Amari, Koji Kurata, and Hiroshi Nagaoka. Information geometry of boltzmann machines. *Neural Networks, IEEE Transactions on*, 3(2):260–271, 1992.

A. P. Dawid, S. Lauritzen, and M. Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012.

H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

I. J. Good. Comment on "measuring information and uncertainty," by R. J. Buehler. In V. P. Godambe and D. A. Sprott, editors, *Foundations of Statistical Inference*, page 337339, Toronto: Holt, Rinehart and Winston, 1971.

M. Gutmann and J. Hirayama. Bregman divergence as general framework to estimate unnormalized statistical models. *arXiv preprint arXiv:1202.3727*, 2012.

G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2002.

G. E Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9 (1):926, 2010.

G. E. Hinton and R. R. Salakhutdinov. A better way to pretrain deep Boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2012.

G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. *MIT Press, Cambridge, Mass*, 1:282–317, 1986.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2005.

A. Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

T. Kanamori and H. Fujisawa. Robust estimation under heavy contamination using unnormalized models. *Biometrika*, 102(3):559–572, 2015.

T. Kanamori. Efficiency bound of local z-estimators on discrete sample spaces. *Entropy*, 18 (7):273, 2016.

K. V. Mardia, J. T. Kent, G. Hughes, and C. C. Taylor. Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika*, 96(4):975–982, 2009.

M. Opper and D. Saad, editors. *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, MA, 2001.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

P. Ravikumar, M. J. Wainwright, J. D. Lafferty. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

T. J. Sejnowski. Higher-order Boltzmann machines. In *American Institute of Physics Conference Series*, volume 151, pages 398–403, 1986.

T. Takenouchi. A novel parameter estimation method for boltzmann machines. *Neural computation*, 27(11):2423–2446, 2015.

T. Takenouchi and T. Kanamori. Empirical localization of homogeneous divergences on discrete sample spaces. In *Advances in Neural Information Processing Systems*, pages 820–828, 2015.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.